We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

4,800
Open access books available

122,000
International authors and editors

135M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

CLARIVATE ANALYTICS BOOK CITATION INDEX INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

Chapter 1

Genome-Guided Transcriptomics, DNA-Protein Interactions, and Variant Calling

Emmanouil E. Malandrakis and Olga Dadali

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.76842

Abstract

Nowadays, molecular biology has definitely become an interdisciplinary science. Toward the study of the functions and the interactions of the biological molecules, such as nucleic acids and proteins, computer science and engineering, along with chemistry and statistics, are routinely engaged. In molecular biology, techniques and methods are constantly developed, and new techniques emerge. Next-generation sequencing and bioinformatics have become the cornerstones of molecular biology. The developing technologies have led to a decrease of the cost per molecular unit analyzed, but at the cost of computer integration and intensification. Many research methods require a reference nucleic acid sequence. Considering the necessary integration of sequencing data and methodology, combining the "omics" approaches can help to elucidate more complex null hypotheses. Here, data processing basics, with an emphasis to commonly used techniques, are summarized. The knowledge gaps are discussed as well as further prospective for integrating next-generation sequencing data.

Keywords: next-generation sequencing, data analysis, Unix, scripting

1. Introduction

The study of the functions and the interaction of the biological molecules such as nucleic acids and proteins has become a daily laboratory routine. By recently, Sanger sequencing was extensively used to uncover new genomic sequences. Sanger sequencing method was named after Frederick Sanger (1918–2013), the British biochemist who invented it and won the Nobel Prize in Chemistry for the second time (1980). Until now, the method is based on PCR amplification and capillary electrophoresis. Each sequencing reaction generates a ladder of ddNTP-terminated, dye-labeled products, which then are submitted to high-resolution electrophoretic

separation within one of 96 or 384 capillaries in one run of a sequencing instrument. The generated fragments are labeled with fluorescent substances and pass the laser, which allows the four different nucleotides to excite and emit different colors of the light spectrum. A camera then captures the colors, and the results are extracted in various formats for further analysis. The analysis of Sanger sequencing data is more or less a straightforward procedure. The sequences can be optically validated and cross-checked through the chromatogram. High-quality reads should not contain ambiguities, and the peaks must be well spaced. On the other hand, poor quality reads have low signal/noise ratio, overlapping peaks and low confidence score. Consequently, a comparison of our sequence can be done with Basic Local Alignment Search Tool (BLAST) by NCBI. BLAST is the cornerstone of sequence analysis, since it facilitates the comparison among amino acid or nucleotide sequences.

Next-generation sequencing (NGS) is an emerging technology with high-throughput outcome. Recently, the rapid development of high-throughput technologies has led to advances in the study of genome function. High-throughput molecular techniques are generally used to study nucleic acids of different species such as DNA, mRNA, lncRNA, etc. Typically, the nucleic acids are fragmented, amplified (or not), and sequenced using various technologies. Although nucleic acid extraction and sequencing are a typical workflow for many laboratories worldwide, integrative software to deal with the analysis workflow in a user-friendly manner is scarce. An intermediate user who is thinking about starting a new NGS project has to set up a Unix-based server with the appropriate software toolkit in order to deal with a huge amount of data. A basic knowledge of R programming language is essential, and the bioconductor project includes an important amount of applications for NGS data processing [1]. Moreover, essential knowledge of scripting languages (Perl, Python) is necessary. In a few words, the data resulting from the sequencer have to be quality checked, filtered, and finally evaluated. This can be achieved on a substantial bioinformatic level.

2. Transcriptomics

Transcriptome sequencing (transcriptomics) enables the characterization of all RNA transcripts for a given organism, including both the coding mRNA and noncoding RNA. For many years, our knowledge on the transcriptome was derived from cloning and sequencing of individual cDNA sequencing. Therefore, it was limited, low-throughput, and partial. However, transcriptomics with next-generation sequencing (NGS) and RNA-Seq is able to increase our knowledge on the dynamic RNA landscape. Compared to the limited capability of Sanger sequencing, a typical RNA-Seq experiment can provide an integrated snapshot of an organism's transcriptome. Normally, regarding RNA-Seq, there are two major experimental setups: de novo assembly of the transcriptome and reference-guided assembly. The former is adequate either when reference genome (or transcriptome) is not available or we want to expand the existing knowledge of an organism's transcriptome. Furthermore, it is mainly utilized in cancer transcriptomics to find fused transcripts or in organisms with trans-splicing. A typical analysis pipeline is presented in **Figure 1**. mRNA sequencing has many advantages over conventional methods. Gene expression can be accurately quantified, and genes with alternative

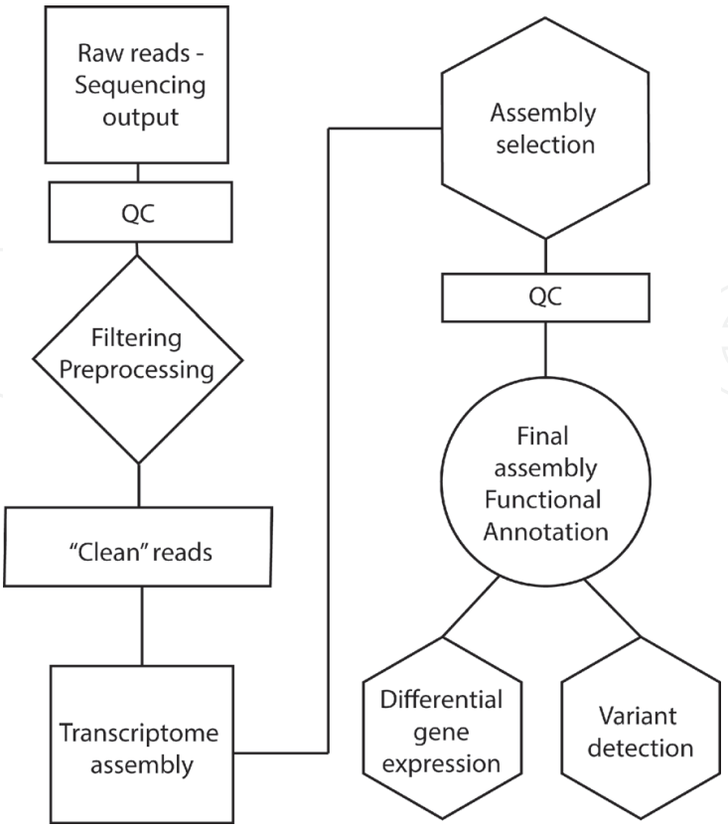

Figure 1. Typical workflow of an RNA-Seq experiment.

splice variants can be identified. Furthermore, reconstructing a transcriptome from short reads is really challenging in terms of computer resources.

To ensure a high-quality transcriptome assembly, the design of the experiment should be carefully designed. If differential expression analysis is planned, biological replication is vital. As a rule of thumb, three to four biological replicates are adequate, but it depends upon the specific experiments. Technical replicates are not essential, but can be used to check for any barcoding effects on results. Usually, technical replicates are highly reproducible (e.g., [2]). In Illumina platforms multiplexing samples are useful for two reasons. Firstly, if a lane fails to produce data, it is still likely that many results from the other lanes can be extracted. Secondly, technical replicates are produced and barcoding effect can be determined.

The very first step upon the receipt of the sequenced data is a secure backup. No one wants to lose precious samples and hundreds of man-hours due to a failure of a hard disk. Data can be stored (and published) in public databases such as NCBI SRA. The reads can be recovered from the database, and a set of metadata is available detailing the experimental conditions.

2.1. Material

A typical flowchart of RNA-Seq experiments usually includes RNA extraction, which must be of high purity and integrity. Most sequencing companies recommend an RNA integrity number (RIN), using Agilent Bioanalyzer 2100 higher than 8. Except for difficult materials (i.e.,

FFPE samples, fossils), normally preserved samples could easily achieve this score, with standard extraction protocols. Two types of protocols are used for RNA extraction: affinity based (in column) and organic extraction (phenol, chloroform, isoamyl alcohol). The former is compatible with various sample types (animal tissue, plant cells, bacteria, yeast, etc.). Furthermore, DNAse treatment which eliminates contaminating genomic DNA is highly facilitated in a column-based extraction. In this way, excellent RNA purity and integrity are achieved. In addition, automated RNA extraction process is able to reduce working time and at the same time provides opportunity in increasing reproducibility and quality of results [3]. Starting from total RNA, two strategies are available for RNA-Seq: enrichment for mature transcripts using poly(A) tails and depletion of abundant ribosomal sequences. In that way, mature mRNA is abundant in the sample for further processing. Since rRNA represents the 80% of the total RNA and mRNA is 5%, mRNA enrichment is crucial in order to achieve a decent sequencing depth. Sequencing depth is the mean number of times that each nucleotide is sequenced. This stands only for genome, where nucleotides remain relatively stable. For transcriptome, differential expression plus biases in sample processing and sequencing can result in genes with lack of coverage.

2.2. Data quality control and filtering

There are numerous pipelines that check the quality of the data produced by the sequencer. Although millions of reads are typically produced by high-throughput sequencers, simple quality controls are essential in order to be sure that the data could be further processed. If any problems or biases are spotted in the dataset, corrective measures can be taken in most cases. FastQC [4] is a very fast and reliable application that can process different data formats such as *fastq*, compressed *fastq*, *SAM*, and *BAM*. By using simple bash scripts, one could easily analyze multiple datasets at once. FastQC can run in a graphical user interface (GUI) environments even in Unix platforms. An application designed to better group FASTQC result data in whole experiments is FQC [5]. The results are stored in simple *html* files and can be viewed with any web browser available. The output includes simple statistics such as the number of reads, sequence length, etc. A more important statistic for the quality of the available reads is the diagram of the quality score over the nucleotide position in the sequence. In the *fastq* format, each read is tagged with a quality score known as Phred quality score. In general, a Phred quality score of 10 means that there is a possibility of the called base being correct of 90%, 20 is 99%, 30 is 99.9%, etc. As a rule of thumb, bases with score over 20 are considered as bases of good quality.

RNA-Seq reads need further preprocessing before assembly and gene expression analysis. Usually, 5′ or 3′ ends present lower-quality or ambiguous sequences. Consequently, these reads are trimmed at both ends. In case the reads have more low-quality or ambiguous nucleotides, they are totally excluded from the analysis. Some good tools for the preprocessing of data include PRINSEQ [6] and Trimmomatic [7]. Although rRNA is routinely removed during library preparation, many sequences are present in raw reads. An efficient tool for rRNA removal is SortMeRNA [8]. SortMeRNA leverages public rRNA databases such as SILVA [9] and Greengenes [10], to identify rRNA sequences. Firstly, developed for metagenomic studies,

the software has the ability to extract rRNA sequences in fastq files for further processing. Although designed for single fastq files, two scripts that split and merge back paired-end files, respectively, are provided. Another one critical step during sequence filtering is adapter and primer removal from the dataset. One could combine adapter clipping with quality trimming with Trimmomatic. Cutadapt [11] is definitely a dedicated tool for adapter and PCR primer clipping and removing. The software supports flexible scanning and removal of contaminating sequences.

Consequently, following the previous steps, the data could be used for further processing. In each step, a quality control could be adequate to safeguard the efficiency of process. As a result, the user is able to further process the data or repeat the step, with different settings, until the results are satisfying.

2.3. De novo transcriptome versus genome-guided assembly

De novo transcriptome assembly is a computer-intensive process. Despite the constant increase of available tools, transcriptome assembly from short reads still remains a very challenging process. Probably, the most popular tool for transcriptome assembly is Trinity [12]. Beyond assembly, Trinity incorporates many post-assembly tools which include assembly QC, full-length transcript analysis, abundance, and differential gene expression analysis. Furthermore, protein-coding analysis and functional annotation software are included.

Evaluating a de novo transcriptome assembly is really a hard job. There is a plethora of metrics to assess the accuracy and completeness of a transcriptome assembly. Honaas et al. [13] concluded that a *combination* of metrics can be used in the following order: a number of reads mapping to the assembly; recovery of conserved, widely expressed genes; N_{50} length statistics; and the total number of unigenes. A number of tools are available for this purpose such as BUSCO [14], DETONATE [15], and TransRate [16].

On the other hand, reference-guided transcriptome assembly could be very solid. However, the accuracy of reference-based transcriptome assembly depends on correct read alignment and genetic variants such as alternative splice variants, CNVs, etc. Transcripts are distinguished from the reference genome by Cufflinks [17], and supporting applications in the suite can be used for further analysis.

2.4. Read mapping and counting

The first major data processing step in sequencing studies for species with a reference genome is the mapping of sequencing reads to the reference (genome or transcriptome). Mapping of the reads is defined as the prediction of the loci from which the reads originate. There are many alignment algorithms such as BWA [18] and Bowtie [19] which are unspliced read mappers and TopHat [20] which is a spliced one. The choice of aligner often influences the final results, as different algorithms show various false-positive and/or false-negative rates. There is no single mapper that can align all reads to a reference. This could be due to sequencing errors or polymorphic loci in the reference. Indeed, unmapped reads could be analyzed for identification of such variants. After mapping, a consequential SAM (Sequence Alignment/Map) file

result is typically converted to the compressed binary version BAM. This is typically achieved with the *samtools view* command [21].

The differential expression analysis of NGS data includes the counting of the mapped reads on the transcripts. The Trinity suite incorporates various applications for further analysis of transcriptome data. Four tools are employed for read counting, namely, alignment-based tools RSEM [22] and eXpress [23], as well as alignment-free kallisto [24] and salmon [25].

2.5. Differential expression analysis

The output from read counting is a matrix of raw counts that is used as input for R-based software such as DESeq [26], DESeq2 [27], or edgeR [28]. Since each software draws upon different statistical methods, differences in outputs may arise. Furthermore, there are available capabilities of clustering the differentially expressed genes or plotting the results in diagrams. One of the most favorite plots available is the heat map, where the expression of the genes is presented in colors that distinguish the control from the treatment groups. Typically, a palette of red and blue colors is used to indicate up- and downregulation, respectively. Finally, the most important thing on the whole procedure is the discovery of genes (or gene clusters) associated with the biological questions posed.

3. DNA-protein interactions

3.1. Introduction

Proteins bind DNA in order to regulate genome function. Among the proteins that bind DNA, most characteristics are the transcription factors (TFs). Transcription factors regulate transcription by switching on and off genes. They act either alone or synergistically with other proteins as cofactors. Furthermore, groups of TFs function in a coordinated fashion to trigger many fundamental genomic processes (cell division, cell death, development) and periodically in reaction to signals coming from outside the cell.

DNA-protein interactions can be studied by using chromatin immunoprecipitation followed by sequencing (ChIP-Seq) [29]. Accordingly, RNA-protein interactions can be unveiled using cross-linking immunoprecipitation (CLIP), RNA-DNA interactions using CHART and CHiRP, and DNA-DNA interactions (using 3C-based methods, including circularized chromosome conformation capture (4C), chromosome conformation capture carbon copy (5C), Hi-C, and chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) [30]. Further down, we are going to focus on DNA-protein interaction methodology. During this step, our purpose is to capture characteristic read distribution at the chromatin interaction sites and detect significantly enriched regions.

3.2. Material

When conducting an immunoprecipitation experiment, probably the most major consideration is the selection of the antibody. The antibody must be specific and work in chromatin

immunoprecipitation. Both monoclonal and polyclonal antibodies are able to work with ChIP, though monoclonal is usually more specific. On the other hand, polyclonal antibodies recognize multiple epitopes on the targets.

This kind of assays starts with cross-linking DNA and protein. During this procedure, the cross-linking substance penetrates intact cells and fixes DNA-protein complexes. The most common stabilizer in ChIP is formaldehyde, and after stabilization chromatin is sheared in fragments commonly 200–600 bp [31].

3.3. Methodology

Following sequencing the dataset has to be aligned on the reference. Typical aligners are used such as Bowtie or BWA, and the corresponding SAM files are converted to their binary analogs (BAM). The alignments can be visualized with the stand-alone genome browser IGV [32]. When uploading an alignment file to the browser, the browser is going to search for the appropriate index file. To create the index file, the BAM file must be sorted according to its chromosomal coordinates. The indexing can be achieved with samtools index.

Another way to visualize these alignments is through the BigWig format which is an indexed binary format. Firstly, BAM is converted into a bedGraph file with BEDTools [33] and then is turned into BigWig using the bedGraphtoBigWig application from the UCSC tools [34]. BEDTools include ready-to-use files for human and mouse genomes. These files can be loaded in genomic viewers such as IGV and zoom in specific genes or chromosomal loci of interest.

MACS analysis [35] was first developed to identify transcription factor-binding sites. MACS empirically models the length of the sequenced ChIP fragments and uses it to improve the spatial resolution of predicted binding sites. MACS can be used for CHIP-Seq data alone or with control sample to increase specificity. In that sense, control sample is highly recommended to distinguish positive binding sites over background noise. Peak files generated from MACS can be uploaded to Ensembl for further analysis. Furthermore, it is advised to look at genes or regulatory elements that are located in proximity with identified regions. PeakAnalyzer [36] is a stand-alone program for the processing of genomic loci, with an emphasis on datasets consisting of ChIP-derived signal peaks. Gene ontology functional annotation can be applied in the closest downstream genes. Finally, it is really interesting to associate motif-binding sites with motifs or sequence patterns. These motifs can be compared to known motifs available in databases such as JASPAR and UniPROBE.

4. Variant calling

4.1. Methodology

Polymorphisms are generally studied in biology, under the prism of various null hypotheses. In population studies, genotype-trait associations, rare diseases and evolutionary biology, and polymorphisms are studied to answer fundamental biological questions. The starting material could be either DNA or RNA, depending on the experimental design. High-quality reads, high

coverage, and a thorough bioinformatic pipelines are prerequisites to identify polymorphisms on a solid basis.

Variant callers demand different preprocessing steps before the actual processing of insertions-deletions (InDels), single-nucleotide polymorphisms (SNPs), and structural variants (SVs). Typical steps include duplicate removals and local realignment. Picard tools by Broad Institute include a Java-based set of command-line tool, including *MarkDuplicates.jar* command for the removal of the duplicate reads. The documentation of the software provides an extended walk-through and describes the metrics produced by the software.

Samtools mpileup is one of the options considered for variant calling. This option demands a reference (either genome or transcriptome) in a FASTA file and the BAM file of the aligned reads. The output is in VCF (variant call format) which is converted to its binary analogue (BCF). Samtools scripts are available that can be used to filter for low mapping quality, low coverage, gaps, and similar biases. All these artifacts are known to increase false-positive rates in SNP calls. VCFtools software [37] has the ability to filter, merge, subset, and query VCF files. Furthermore, it is able to produce simple descriptive statistics such as InDel length, transversion/transition ratio, etc. All these results can be visualized either with simple tools such as the *tview* command of the samtools package or with more sophisticated viewers such as IGV [32].

4.2. Annotation

Raw variant calling files contain many false-positive results, which may be due to the sequence quality of the reads, PCR artifacts, or other biases. Annotating these variations may mark these SNVs as less confident. In addition, important mutations could be identified according to the effect they bear on the genome. For these purposes, two potential applications are Annovar [38] and SnpEff [39]. While the latter is able to use directly VCF files, Annovar uses a specific input format that files should be converted to.

Using predefined gtf (general transfer format) models, all SNPs can be classified as synonymous, non-synonymous, loss of function, start loss, stop loss, start gain, start loss, etc. according to their effect on the genome. The 1000 Genomes data as well as the dbSNP can be used to extract data of features for annotated genomes. In case of non-model species, genes should firstly be dully annotated. Finally, any important variants should be spotted using Sanger sequencing for validation.

5. Perspectives

Although thousands of papers have been published, many things have to be done toward integration of NGS data and processing. Most of the work done is not reproducible, and data processing pipelines could not be shared among different experiments. Although guidelines for result validation have been extensively reviewed [40–43], processing parameters should be recorded thoroughly. The complexity of the NGS experiments demands complete description

of the parameters through metadata; minimum information about any (x) sequence (MIxS) creates a single-entry point to all minimum information checklists for sequence data [44].

Beside reproducibility issues, one could safely argue that NGS data processing is not actually user-friendly. Expertise in informatics and more specifically in Unix-based systems is essential. In that sense, biologists are able to handle sequencing data in association with computer scientists. Furthermore, for assembly and annotation purposes, intense computing is needed that diverges from personal computers' capabilities. Therefore, small servers to large computing clusters need to be employed for processing. The development of graphical user interface (GUI) software for NGS processing is essential. Furthermore, software suites that include all steps of processing (QC, preprocessing, filtering, assembling, mapping, and differential analysis) combined with machine learning systems could facilitate analysis from beginners or intermediate computer users. In other words, more sophisticated software could propose the user, according to the experiment and the data walk-through to analyze the dataset.

Author details

Emmanouil E. Malandrakis* and Olga Dadali

*Address all correspondence to: emalandrak@uth.gr

University of Thessaly, Volos, Greece

References

[1] Gentleman RC et al. Bioconductor: Open software development for computational biology and bioinformatics. Genome Biology. 2004;**5**(10):R80

[2] Mortazavi A et al. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature Methods. 2008;**5**(7):621-628

[3] Tan SC, Yiap BC. DNA, RNA, and protein extraction: The past and the present. Journal of Biomedicine & Biotechnology. 2009;**2009**:574398

[4] Andrews S. FastQC A Quality Control tool for High Throughput Sequence Data. Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

[5] Brown J, Pirrung M, McCue LA. FQC Dashboard: integrates FastQC results into a web-based, interactive, and extensible FASTQ quality control tool. Bioinformatics. 2017;**33**(19): 3137-3139

[6] Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. Bioinformatics. 2011;**27**(6):863-864

[7] Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics. 2014;**30**(15):2114-2120

[8] Kopylova E, Noe L, Touzet H. SortMeRNA: Fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics. 2012;**28**(24):3211-3217

[9] Yilmaz P et al. The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. Nucleic Acids Research. 2014;**42**(Database issue):D643-D648

[10] DeSantis TZ et al. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Applied and Environmental Microbiology. 2006;**72**(7):5069-5072

[11] Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011;**17**(1):10-12

[12] Haas BJ et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature Protocols. 2013;**8**(8):1494-1512

[13] Honaas LA et al. Selecting superior de novo transcriptome assemblies: Lessons learned by leveraging the best plant genome. PLoS One. 2016;**11**(1):e0146062

[14] Simao FA et al. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;**31**(19):3210-3212

[15] Li B et al. Evaluation of de novo transcriptome assemblies from RNA-Seq data. Genome Biology. 2014;**15**(12):553

[16] Smith-Unna R et al. TransRate: Reference-free quality assessment of de novo transcriptome assemblies. Genome Research. 2016;**26**(8):1134-1144

[17] Trapnell C et al. Differential analysis of gene regulation at transcript resolution with RNA-seq. Nature Biotechnology. 2013;**31**(1):46-53

[18] Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;**25**(14):1754-1760

[19] Langmead B et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology. 2009;**10**(3):R25

[20] Trapnell C, Pachter L, Salzberg SL. TopHat: Dscovering splice junctions with RNA-Seq. Bioinformatics. 2009;**25**(9):1105-1111

[21] Li H et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; **25**(16):2078-2079

[22] Li B, Dewey CN. RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011;**12**:323

[23] Roberts A, Pachter L. Streaming fragment assignment for real-time analysis of sequencing experiments. Nature Methods. 2013;**10**(1):71-73

[24] Bray NL et al. Near-optimal probabilistic RNA-seq quantification. Nature Biotechnology. 2016;**34**(5):525-527

[25] Patro R et al. Salmon provides fast and bias-aware quantification of transcript expression. Nature Methods. 2017;**14**(4):417-419

[26] Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biology. 2010;**11**(10):R106

[27] Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology. 2014;**15**(12):550

[28] Robinson MD, McCarthy DJ, Smyth GK. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;**26**(1):139-140

[29] Solomon MJ, Larsen PL, Varshavsky A. Mapping protein DNA interactions in vivo with formaldehyde – Evidence that histone-H4 is retained on a highly transcribed gene. Cell. 1988;**53**(6):937-947

[30] Sims D et al. Sequencing depth and coverage: Key considerations in genomic analyses. Nature Reviews. Genetics. 2014;**15**(2):121-132

[31] Park PJ. ChIP-seq: Advantages and challenges of a maturing technology. Nature Reviews. Genetics. 2009;**10**(10):669-680

[32] Robinson JT et al. Integrative genomics viewer. Nature Biotechnology. 2011;**29**(1):24-26

[33] Quinlan AR, Hall IM. BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;**26**(6):841-842

[34] Kuhn RM, Haussler D, Kent WJ. The UCSC genome browser and associated tools. Briefings in Bioinformatics. 2013;**14**(2):144-161

[35] Zhang Y et al. Model-based analysis of ChIP-Seq (MACS). Genome Biology. 2008;**9**(9): R137

[36] Salmon-Divon M et al. PeakAnalyzer: Genome-wide annotation of chromatin binding and modification loci. BMC Bioinformatics. 2010;**11**:415

[37] Danecek P et al. The variant call format and VCFtools. Bioinformatics. 2011;**27**(15):2156-2158

[38] Wang K, Li M, Hakonarson H. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Research. 2010;**38**(16):e164

[39] Cingolani P et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin). 2012;**6**(2):80-92

[40] Roy S et al. Standards and guidelines for validating next-generation sequencing bioinformatics pipelines: a joint recommendation of the association for molecular pathology and the college of american pathologists. The Journal of Molecular Diagnostics. 2018;**20**(1):4-27

[41] Jennings LJ et al. Guidelines for validation of next-generation sequencing-based oncology panels: A joint consensus recommendation of the association for molecular pathology and college of american pathologists. The Journal of Molecular Diagnostics. 2017;**19**(3):341-365

[42] Kim J et al. Good laboratory standards for clinical next-generation sequencing cancer panel tests. Journal of Pathology and Translational Medicine. 2017;**51**(3):191-204

[43] Endrullat C et al. Standardization and quality management in next-generation sequencing. Applied & Translational Genomics. 2016;**10**:2-9

[44] Yilmaz P et al. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. Nature Biotechnology. 2011;**29**:415